Comparison of longitudinal CA125 algorithms as a first line screen for ovarian cancer in the general population

Oleg Blyuss[1], Matthew Burnell[1], Andy Ryan[1], Aleksandra Gentry-Maharaj[1], Inés P. Mariño[1,2], Jatinderpal Kalsi[1], Ranjit Manchanda[1,6], John F. Timms[1], Mahesh Parmar[3], Steven J. Skates[4], Ian Jacobs[1,5,8], Alexey Zaikin[1,7*], and Usha Menon[1*].
* joint

[1] Women's Cancer, Institute for Women's Health, University College London, Gower Street, London, WC1E 6BT, UK;
[2] Departamento de Biología y Geología, Física y Química Inorgánica, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain;
[3] Medical Research Council Clinical Trials Unit at UCL, London, UK
[4] Massachusetts General Hospital, Boston, MA, USA;
[5] Faculty of Medical and Human Sciences, 1.018 Core Technology Facility, University of Manchester, Grafton Street, M13 9NT, UK;
[6] Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ
[7] Department of Mathematics, University College London, Gower Street, London, WC1H 0AY, UK;
[8] Faculty of Medicine, UNSW Sydney, Kensington High Street, Sydney, NSW 2052, Australia

Corresponding Author
Professor Usha Menon
Gynaecological Cancer Research Centre, Department of Women's Cancer
Institute for Women's Health
1st Floor Maple House, 149 Tottenham Court Road
London W1T 7DN
Tel: 0203 447 2108; email: u.menon@ucl.ac.uk

Keywords
UKCTOCS, CA125, Ovarian cancer, screening, longitudinal algorithms, PEB, MMT, ROCA

Conflicts of interest

UM has stock ownership and has received research funding from Abcodia. She has received grants from the Medical Research Council (MRC), Cancer Research UK, (CR UK) the National Institute for Health Research (NIHR), and The Eve Appeal. IJJ reports personal fees from and stock ownership in Abcodia as the non-executive director and consultant. He reports personal fees from Women's Health Specialists as the director. He has a patent for the Risk of Ovarian Cancer algorithm and an institutional license to Abcodia with royalty agreement. He is a trustee (2012–14) and Emeritus Trustee (2015 to present) for The Eve Appeal. He has received grants from the MRC, CR UK, NIHR and The Eve Appeal. SJS reports personal fees from the LUNGevity Foundation and SISCAPA Assay Technologies as a member of their Scientific Advisory Boards. He reports personal fees from Abcodia as a consultant and AstraZeneca as a speaker honorarium. He has a patent for the Risk of Ovarian Cancer algorithm and an institutional license to Abcodia. All other authors declare no competing interests.

Abstract

Purpose In the United Kingdom Collaborative Trial of Ovarian Cancer Screening(UKCTOCS) women in the multimodal (MMS) arm had a serum CA125 test (first-line), with those at increased risk, having repeat CA125/ultrasound (second-line test). CA125 was interpreted using the 'Risk of Ovarian Cancer Algorithm'(ROCA). We report on performance of other serial algorithms and a single CA125 threshold as a first line screen in the UKCTOCS dataset.

Experimental Design 50,083 post-menopausal women who attended 346,806 MMS screens were randomly split into training and validation sets, following stratification into cases (ovarian/tubal/peritoneal cancers) and controls. The two longitudinal algorithms, a new serial algorithm, method of mean trends (MMT) and the parametric empirical Bayes (PEB) were tested run in the blinded validation set and the performance characteristics, including that of a single CA125 threshold, were compared.

Results The area under receiver operator curve (AUC) was significantly higher (p=0.01) for MMT (0.921) compared to CA125 single threshold (0.884). At a specificity of 89.5%, sensitivities for MMT (86.5%; 95%CI:78.4-91.9) and PEB (88.5%; 95%CI: 80.6-93.4) were similar to that reported for ROCA (sensitivity 87.1%; specificity 87.6%; AUC 0.915) and significantly higher than the single CA125 threshold (73.1%; 95%CI: 63.6-80.8).

Conclusions These findings from the largest available serial CA125 data set in the general population provide definitive evidence that longitudinal algorithms are significantly superior to simple cut-offs for ovarian cancer screening. Use of these newer algorithms requires incorporation into a multimodal strategy. The results highlight the importance of incorporating serial change in biomarker levels in cancer screening/early detection strategies.

Translational Relevance

Earlier diagnosis of ovarian cancer remains a key need as it continues to be the leading cause of death from gynecological cancer, accounting for 5% of all female cancer deaths. In the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) multimodal screening incorporating a longitudinal CA125 algorithm (ROCA) outperformed threshold rules and resulted in significant detection of earlier stage disease but no definitive mortality reduction compared to no screening. We now show that other longitudinal CA125 algorithms, a newly developed, Method of Mean Trends (MMT), and the published Parametric Empirical Bayes (PEB) algorithm have comparable performance as a first line annual test and also significantly outperform simple cut-offs. The advantages of these new algorithms are computational simplicity with incorporation of additional biomarkers much easier. These findings highlight the need to incorporate serial change in biomarker levels for screening/early detection of cancer. While ovarian cancer screening is not recommended in the general population, our findings have immediate implications for high-risk women in countries where twice a year CA125 screening is an option. It highlights the importance to look at trends and not absolute cut-offs alone. Use of the new algorithms requires incorporation into a multimodal strategy and evaluation in clinical trials to assess overall performance**.**

Introduction

Ovarian cancer remains the leading cause of death from gynecological cancer among women and accounts for 5% of all female deaths from cancer, corresponding to annual deaths of around 4,100 in the UK (1), 42,700 in Europe, 22,280 in USA (2) and 152,000 worldwide (1). Most women are diagnosed in advanced stage (Stage III-IV) with reported 5-year survival rates of 19% (Stage III) and 3% (Stage IV) respectively. The higher survival rates of 70-90% in earlier stage (Stage I-II) disease has driven international screening efforts to detect the disease earlier (3). To date the large screening trials have used measurement of a tumour marker CA125 (Cancer Antigen 125 protein) in the blood and transvaginal ultrasound to image the ovaries.

In the ovarian component of the Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), an absolute CA125 cut-off of 35 U/mL and pelvic ultrasound was used as first line annual tests. There was no stage shift or reduction in ovarian cancer deaths between the screen and no screening (control) arms (4). More recently in UKCTOCS, multimodal screening (MMS) resulted in significant detection of earlier stage disease in women with invasive epithelial ovarian/tubal/peritoneal cancers (iEOC/PPC) compared to the control arm. In a pre-specified subgroup of the primary mortality analysis ROCA reduced mortality in the 80% of cancers where a baseline CA125 was measured, that is, in incident cancers. However further follow-up is needed to assess whether screening results in a definitive mortality reduction for all ovarian cancers (5). In MMS the annual first-line test was CA125 which was interpreted using the longitudinal Risk of Ovarian Cancer Algorithm (ROCA). Women found to be at increased risk had repeat CA125 and/or transvaginal ultrasound (second-line test).

As a first line test, ROCA had significantly better performance characteristics for detection of invasive epithelial ovarian/tubal cancer (iEOC/PPC) compared to a CA125 cut-off in the UKCTOCS dataset. Only 48% of the incident cases would have been detected at the last annual incidence screen if a CA125 cut-off of 35 U/mL had been used (6). The statistical model underlying the ROCA is built on two important assumptions: (i) each woman has her own baseline CA125 level and variation about this level, and (ii) after cancer inception the tumor sheds CA125 into the circulation whereupon serum CA125 increases exponentially reflecting tumor doubling. ROCA best detects cancers where a significant increase above a woman's CA125 baseline occurs; hence the pre-specified subgroup analysis for cancers where a CA125 baseline was measured during screening. The second assumption corresponds to a change point in the serum CA125 time series as the cancer develops (7-9). The development of new ovarian cancer detection algorithms that further minimize assumptions remains an important scientific goal as does the performance of serial algorithms compared to single thresholds in the context of screening. Parametric empirical Bayes (PEB) (10) is another algorithm that has been described for interpreting serial CA125 data.

In this paper we use the data from 50,083 post-menopausal women who underwent 346,806 annual screens and follow-up in the multimodal (MMS) arm during the course of UKCTOCS to (1) build a new algorithm for longitudinal analysis of cancer biomarkers, "method of mean trends" (MMT), which measures the dynamics of the biomarker over time using multiple trend indices and (2) investigate the performance of both longitudinal biomarker algorithms (MMT and PEB) and CA125 cut-off as first line tests for ovarian cancer screening.

Patients and methods

In UKCTOCS, 202,638 low-risk postmenopausal women aged 50-74 were randomized between 2001 and 2005 to one of two screening (ultrasound: USS; multimodal: MMS) or a no screening (control: C) arm in a ratio of 1:1:2. Exclusion criteria were self-reported previous bilateral oophorectomy or ovarian malignancy, increased risk of familial ovarian cancer, or active non-ovarian malignancy. The trial was approved by the United Kingdom North West Multicentre Research Ethics Committee (ISRCTN22488978) and listed on ClinicalTrials.gov (NCT00058032). Trial design, including details of recruitment and randomization have been described elsewhere (5, 6, 11). All women provided written informed consent.

Women were offered annual screening from randomization to 31st December 2011. The screening protocol and management of screen-detected abnormalities have been previously described (6, 11). In brief, in the MMS group women had an annual serum CA125 which was interpreted using ROCA. If the 'risk of ovarian cancer' was normal they were triaged to annual screening; if intermediate they had repeat CA125 in three months and if elevated, they had repeat CA125 and transvaginal scan.

All volunteers were followed using their National Health Service number through data linkage with the appropriate national agencies for cancer registrations and/or deaths as well as by postal questionnaires. Primary cancer site, morphology, stage and grade were assigned as of 31st December 2014 following review of all medical notes by an independent outcomes review committee (two pathologists and two gynecological oncologists) who were blinded to the randomization group as previously described (5, 6, 11).

Sample set for current analysis

The sample set is derived from data on all eligible women randomized to the MMS arm who were included in the mortality analysis and had a CA125 measurement (5). 'Cases' were women confirmed at censorship (31st December 2014) by the outcomes review committee to have iEOC/PPC, borderline epithelial and non-epithelial ovarian cancer. Controls were all women who did not have ovarian/tubal/peritoneal cancer. The dataset of eligible women was randomly divided in a stratified manner, so that both controls and cases were each split in a 1:1 ratio into a training and validation set.

Since the ROCA was prospectively evaluated in UKCTOCS, following the annual screen, all repeat CA125 tests were triggered by ROCA. Hence to limit concerns over potential bias, the CA125 data for this analysis was limited to annual measurements. All CA125 values were transformed by taking a logarithm of their values prior to applying the algorithms since on this scale the distribution was much closer to a Normal distribution than the original scale.

Development/training of algorithm

***Method of Mean Trends (MMT) algorithm*:** This new method evaluates the dynamics of longitudinal markers by averaging weighted derivatives of marker changes for all intervals of time between measurements. Since the most recent biomarker measurement is more important than all previous ones, weights were proposed in order to take into account the importance of those samples, which were closer to the most recent observation. For each individual woman *"i"* the whole serial pattern, $Y_{i,j}, j = 1..T,$ was mapped into a new five-

variable space including its mean derivative, the three indices described below and the most recent measurement $Y_{i,T}$. In this way, instead of the initial collection of T measurements over time for a particular marker, the dimension to 5 variables for the CA125 marker was reduced. For this marker, and each interval between two consecutive measurements, the derivative was approximated using the expression $\Delta Y_{ij}/\Delta t_{ij}$, where $\Delta Y_{i,j} = Y_{i,j+1} - Y_{i,j}$, $\Delta t_{i,j} = t_{i,j+1} - t_{i,j}$, then the mean derivative was calculated, giving the most recent measurement higher weight $\sum_{j=1}^{T-1} w_{ij} \frac{\Delta Y_{i,j}}{\Delta t_{i,j}}$, where the weights w_{ij} were computed for each interval between sequential samples as:

$$w_{ij} = \frac{1}{t_{i,T} - (t_{i,j+1} + t_{i,j})/2},$$

where $t_{i,T}$ was the age of the patient at the time of the most recent sample while $t_{i,j}$ was the age of the patient when the j-th sample was taken. In this way, more recent measurements were provided a higher weighting. Apart from the mean derivative, multiple indices were analyzed, and after using Akaike Information Criterion (AIC) for the model selection, which deals with the trade-off between the goodness of fit of the model and the simplicity of the model, three further indices were used as additional parameters of the MMT:

$$A_i = \left(\sum_{j=1}^{T-1} \frac{\Delta Y_{i,j} \cdot \Delta t_{i,j}}{2} \right) \Big/ (T-1) \quad (1)$$

$$B_i = \sqrt{\frac{\sum_{j=1}^{T} \left(Y_{i,j} - \overline{Y_i}\right)^2}{T}} \Big/ \overline{Y_i} \quad (2)$$

$$C_i = \frac{\sum_{j=1}^{T} Y_{i,j} t_{i,j}}{\sum_{j=1}^{T} t_{i,j}} \quad (3)$$

As a final step, MMT used a logistic regression model based on the weighted average derivatives, the described indices 1 to 3, and the latest currently available measurement

taken for each patient. This logistic regression model (12) was then fitted to obtain coefficients, which provide predictions on the probability scale which were the basis of the rule for classification into cases and controls. If there was only a single CA125 value, the threshold at 90% specificity was used for classification. An important advantage of this proposed approach is the ability to use more than one marker simultaneously by adding into the logistic regression model-average derivatives and other indices calculated separately for each of marker $m = 1..M$.

In summary, the MMT algorithm applied for the prediction of disease based on serial measurements is as follows:

- *Step 1: approximate the time-derivatives of the biomarker series, for each patient and each measurement, as* $\Delta Y_{ij}/\Delta t_{ij}$
- *Step 2: calculate the weight for each derivative as* $w_{ij} = \frac{1}{t_{i,T} - (t_{i,j+1} + t_{i,j})/2}$
- *Step 3: calculate the weighted mean* $\sum_{j=1}^{T-1} w_{ij} \frac{\Delta Y_{i,j}}{\Delta t_{i,j}}$
- *Step 4: calculate the indices* A_i, B_i, C_i *in expressions (1), (2) and (3)*
- *Step 5: use AIC to select the predictors (out of weighted derivative, indices* A_i, B_i, C_i *and raw measurement of the biomarker) that best explains the labels of the patients (control=0, case=1)*
- *Step 6: fit the logistic regression with the selected predictors and the labels of all patients.*

Once the logistic regression is fitted it can be used to predict the risk of the disease for the new patient. If more than 1 biomarker measurement is available for all the patients, the procedure above is repeated calculating the 5 predictors for each biomarker (Step 1-4) and including them all in the AIC variable selection step (Step 5).

Parametric Empirical Bayes (PEB) algorithm: This method, described previously (10), allows calculation of a biomarker threshold for each subject based on their previous screening history. The approach requires a serial pattern of markers in healthy women and can be used for analysing the performance of new individual markers. Suppose a subject with n historical screens with an average marker concentration $\overline{y_n}$ is going to have a new screen and assuming that we operate at level α, the threshold given by the PEB algorithm is:

$$T = \mu + (\overline{y_n} - \mu)B_n + z_\alpha\sqrt{1 - B_1 B_n}\sqrt{V},$$

where μ is the population mean, $B_n = \frac{\tau^2}{\sigma^2/n+\tau^2}$; $V = Var[Y_{i,j}]$ is the variance of measurements $Y_{i,j}$; $\sigma^2 \; and \; \tau^2$ are the within-subject and between-subject variances, $\sigma^2 = \left(\frac{1}{2}\right) Var\left[Y_{i,2} - Y_{i,1}\right], \tau^2 = V - \sigma^2$; z_α is the α-quantile of a standard normal distribution ($z_\alpha = 1.96$ when $\alpha = 0.975$). At the initial screen $B_0 = 0$ and so the threshold becomes:

$$T = \mu + z_\alpha\sqrt{V}.$$

It should be noted here that, since after obtaining all the parameters including the level α from the training set for each patient the PEB algorithm yields an outcome of 0 (no cancer) or 1 (cancer present), depending on whether the last measurement is higher than the threshold or not, we consider only the value of the sensitivity at a fixed level of specificity in the sequel. The area under the ROC curve cannot be used to analyse performance of the PEB algorithm because in this setting the outcome (0 or 1) does not allow the use of thresholds required for ROC curve construction.

Training set

OB was provided all data including CA125 measurements for each woman, dates of birth, dates when the measurements were taken, case-control status and dates of diagnosis for cancer cases on the training set. The MMT and PEB were developed/trained respectively using the training set by OB. The annual CA125 values were used in a sequential manner. At each annual screen, all previous annual measurements were incorporated and a new PEB/MMT classification was determined. The outcome for the PEB was either 0 or 1 for each measurement in a longitudinal time series while that for MMT was continuous results of the logistic regression. At the training stage a threshold was calculated for the MMT to provide similar specificity to annual ROCA classification of 'normal risk' which was 87.6% in UKCTOCS[6]. Since MMT uses trend indices, which cannot be calculated for the first annual measurement (no previous history), the 0.9 quantile was calculated from the control measurements. After that, for every patient, the first annual measurement was compared to this quantile and depending on whether the measurement was higher or not, risk was assigned as abnormal or normal respectively. Sensitivity, specificity and AUC were calculated.

Validation set

The validation set comprised of a set of women with their serial annual CA125 measurements but no outcomes. OB as described above, normalized all measurements by taking a logarithm of their values. The annual CA125 values were used in a sequential manner. At each annual screen, all previous annual measurements were incorporated and a PEB and MMT classification (and prediction probability) was calculated. The data was then transferred to MB who unblinded the outcome data and compared the performance of the two algorithms and the single CA125 cut-off.

Statistical analysis

The primary outcome was iEOC/PPC diagnosed within 1 year of annual CA125 measurement. Women with borderline epithelial or non-epithelial ovarian cancer were excluded from this analysis. Secondary outcome was all primary ovarian/tubal/peritoneal cancers and the whole data set was used for this analysis. When dealing with the determination of outcomes, the last blood sample was considered as a true positive (if within 1 year from diagnosis) and all prior annual samples as true negatives. A subgroup analysis was undertaken which was restricted to cancers diagnosed between 1-2 years from last measurement in order to see if there was any difference in lead time between the algorithms. For this analysis, if there was more than one measurement within 1-2 years then the closest to 2 years was used as the 'last measurement'. As above the last measurement was considered as true positive and prior annual samples considered true negatives. All annual samples beyond the last measurement were discarded. When dealing with controls, all samples were included as true negatives.

The performance characteristics of the two algorithms and CA125 were evaluated and compared in terms of 1) the sensitivity (proportion detected of those with cancer) at a fixed specificity (proportion of controls correctly detected not to have cancer): for PEB the threshold was implicit in its formulation; for MMT the threshold was the value which provided 0.9 specificity in the training set; for CA125 the threshold was the common value of $\geq$30 U/ml for postmenopausal women and 2) the area under the Receiver Operating Characteristic (ROC) curve (AUC). Inference for the ROC curves was based on cluster-robust standard errors that accounted for the serially correlated nature of the samples. It was not possible to create AUC for PEB given the outcome was not continuous.

Results

The eligible women comprised all 50,083 of the 50,624 randomised to the MMS group who attended for screening and had an annual serum CA125 level measured. They underwent 347,002 annual screens (median 8, IQR 6-9). Median follow-up was 11.1 years (IQR 10.0–12.0).(5)

During follow-up, 332 developed ovarian/tubal/peritoneal cancer as of 31st December 2014. The training set comprised of 25,041 women with 161(139 iEOC/PPC) cases and the validation set 25,042 women with 171(143 iEOC/PPC) cases (Table 1). Baseline characteristics of women included in the training and validation sets were balanced (Table 2). Morphology of cases together with histological subtype and stage of iEOC/PPC are presented in Table 3. Longitudinal algorithms were applied to the validation set, which contained 174,270 annual CA125 measurements from 25,042 women (Table 1).

Table 1: Details of cases and controls in training and validation sets

	Overall		**Primary analysis - cancer diagnosed <1 year of sample**		**Secondary analysis - cancer diagnosed >1 and <2 years after sample**	
	No of women	**No of annual CA125**	**No of women**	**No of annual CA125**	**No of women**	**No of annual CA125**
Primary outcome - Invasive epithelial ovarian/tubal/peritoneal cancer						
Training set						
Cases	139	621	91	375		
Controls	24880	172039	24880	172039		
Validation set						
Cases	143	666	104	466	90	383
Controls	24871	173478	24871	173478	24871	173478
Secondary outcome - Ovarian*/tubal/peritoneal cancers						
Training set						
Cases	161	693	108	433		

Controls	24880	172039	24880	172039		
Validation set						
Cases	171	792	123	553	109	468
Controls	24871	173478	24871	173478	24871	173478

Abbreviations: CA125, cancer antigen 125
* includes borderline, non-epithelial and invasive epithelial

Table 2: Baseline characteristics of cases and controls in training and validation sets

Baseline characteristics	Training set	Validation set
No of women	25041	25042
Median age at recruitment (years)	60.60	60.68
BMI	25.7	25.72
OCP use	59.5%	59.3%
Duration of OCP use (years)	5	5
Hysterectomy	19.4%	19.0%
% White ethnicity	97.0%	97.0%
HRT use	18.7%	18.7%
Personal history of breast cancer	3.7%	3.7%

Abbreviations: BMI, body mass index; OCP, oral contraceptive pill; HRT, hormone replacement therapy

Table 3: Morphology of ovarian cancer cases used for training and validation of the algorithms

Characteristics	Training set (161)	Validation set (171)	Validation set - primary analysis	Validation set - secondary analysis
Morphology of cases				
Non-epithelial ovarian cancer	3	5	3	4
Borderline epithelial ovarian cancer	19	23	16	15
Invasive epithelial ovarian/tubal cancer	135	131	79	93
Primary peritoneal cancer	4	12	11	11
Total	161	171	109	123
Histological type of invasive epithelial ovarian/tubal/peritoneal cancer				
Type I	21	23	13	17
Endometrioid (low grade)	9	6	3	5
Serous (low grade)	6	5	2	5
Clear cell	5	9	8	6
Mucinous	1	3	0	1
Type II	103	113	74	84

High grade serous	81	90	62	68
Carcinoma, NOS	14	8	4	4
Endometrioid (high grade)	6	10	6	9
Carcinosarcoma	2	5	2	3
Type uncertain	15	7	3	3
Carcinoma, NOS	10	6	3	3
Serous (grade unknown)	4	1	0	0
Small cell carcinoma	1	0	0	0
Total	139	143	90	104
Stage of invasive epithelial ovarian/tubal/peritoneal cancer				
I	36	33	19	24
II	11	20	16	17
III	74	76	47	57
IV	18	13	8	6
Total	139	142*	90	104

Abbreviations: NOS, not otherwise specified
*Unable to stage one case in Validation Set
** One woman diagnosed with small cell carcinoma in the training set

Figure 1 shows the ROC curves for MMT and the CA125 threshold rules for detection of iEOC/PPC cases diagnosed within 1 year of last annual sample (primary analysis). MMT provided a higher area under the curve, 0.921 compared with 0.884 for the single threshold rule. The AUC for CA125 single threshold was significantly lower than MMT (p=0.01).

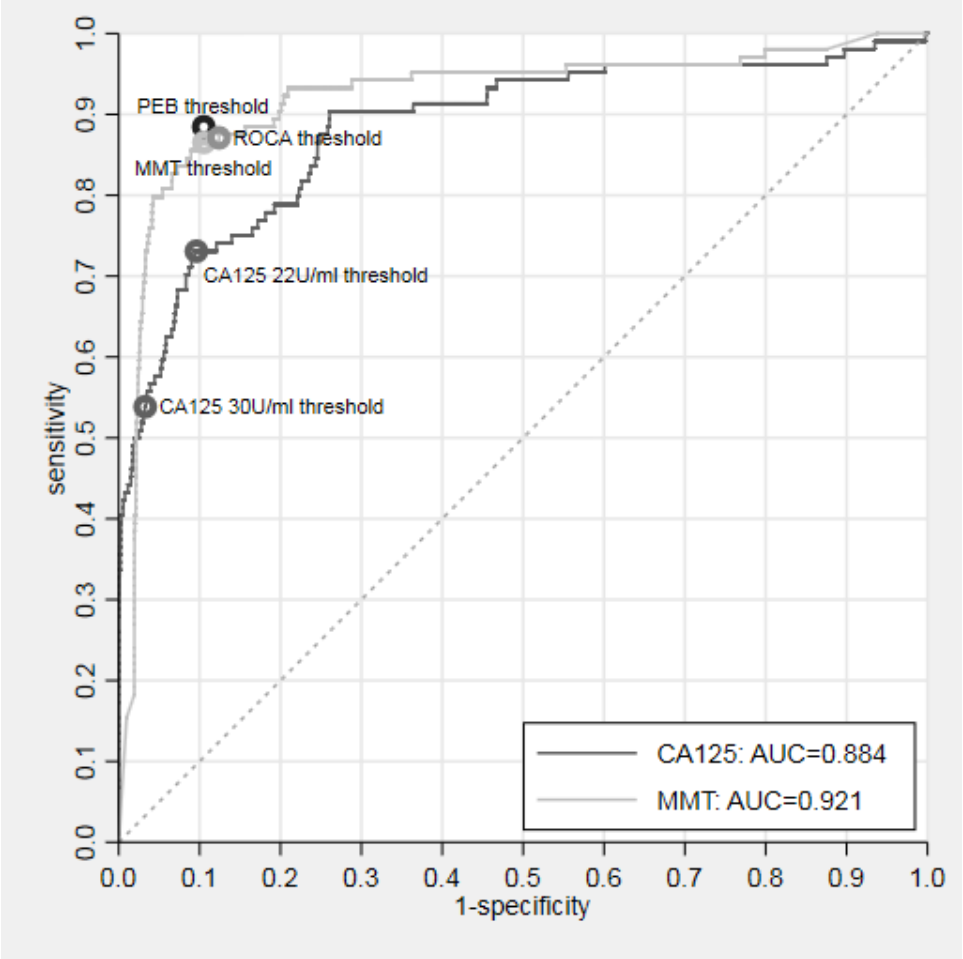


Figure 1 Performance characteristics of of CA125 interpretted using MMT, threshold rules, PEB and ROCA for detection of iEOC/PPC cases. Circle points give particular values of sensitivity and specificity provided by MMT and PEB corresponding to cut-offs obtained from the training set (MMT and PEB), CA125 using 22 and 30 U/ml cut-offs and ROCA as reported in (6) .

Abbreviations: PEB, parametric empirical Bayes; MMT, method of mean trends; CA125, cancer antigen 125; AUC, area under roc-curve

At a specificity of 89.5% for PEB (for which it was not possible to compute AUC), sensitivities were 73.1% (95%CI: 63.6-80.8) for the single CA125 threshold, 86.5% (95%CI: 78.4-91.9) for MMT and 88.5% (95%CI: 80.6-93.4) for PEB. In a hypothetical cohort of 100,000 women with an average incidence of about 50 per 100,000 a year this result would

imply that MMT would detect about (86.5-71.3)*50/100≈7 extra cases and PEB would detect about (88.5-71.3)*50/100≈8 extra cases compared to the CA125 cut-off. To assess the significance of differences in sensitivity at fixed specificity for different algorithms, McNemar's exact test was used. The sensitivity was significantly different compared to the single threshold rule. The longitudinal approaches were not significantly different from each other. 11.5% (12/104) of iEOC/PPC were not detected on the last annual screen by either longitudinal algorithm.

Table 4 shows the sensitivity, specificity and AUC confidence intervals for each of the algorithms and the CA125 single threshold rule in the primary and secondary analyses for the primary and secondary outcomes for both sets. Both longitudinal algorithms provided similar characteristics for both outcomes in the primary and secondary analyses. In all the subgroups of the analysis, PEB and MMT provided higher sensitivity compared with the single CA125 threshold.

Table 4: Cut-point sensitivity and specificity and area under curve (AUC) for primary and secondary analyses

	Primary analysis – cancer diagnosed <1 year after sample			**Secondary analysis - cancer diagnosed 1-2 years after sample**		
	PEB	**MMT****	**CA125 cut-off >30 U/ml**	**PEB**	**MMT****	**CA125 cut-off >30 U/ml**
Primary outcome - Invasive epithelial ovarian/tubal/peritoneal cancer						
Training set						
Sensitivity (% with 95%CI)	85.7 (76.8-92.2)	86.8 (78.1-93.0)	58.2 (47.4-68.5)			
Specificity (% with 95%CI)	90.4 (90.2-90.5)	89.5 (89.3-89.6)	96.7 (96.6-96.8)			
AUC (% with 95%CI)		91.5 (87.8-95.2)	89.6 (85.9-93.3)			
Validation set						
Sensitivity (% with 95%CI)	88.5 (80.6-93.4)	86.5 (78.4-91.9)	53.8 (44.1-63.3)	26.7 (18.4-36.9)	23.3 (15.6-33.4)	8.9 (4.4-17)
Specificity (% with 95%CI)	89.5 (89.3-89.7)	89.5 (89.2-89.7)	96.7 (96.5-96.9)	89.5 (89.3-89.6)	89.4 (89.2-89.7)	96.7 (96.5-96.9)

AUC (% with 95%CI)		92.1 (88.7-95.4)	88.4 (84.4-92.4)		61.3 (55-67.6)	59.8 (53.7-65.8)
Secondary outcome - All Ovarian*/tubal/peritoneal cancers						
Training set						
Sensitivity (% with 95%CI)	82.4 (73.9-89.1)	84.3 (76.0-90.6)	54.6 (44.8-64.2)			
Specificity (% with 95%CI)	90.4 (90.2-90.5)	89.5 (89.3-89.6)	96.7 (96.6-96.8)			
AUC (% with 95%CI)		89.9 (86-93.8)	87.9 (84.3-91.6)			
Validation set						
Sensitivity (% with 95%CI)	85.4 (77.8-90.6)	84.6 (76.9-90)	50.4 (41.5-59.3)	28.4 (20.7-37.8)	24.8 (17.5-33.9)	8.3 (4.3-15.3)
Specificity (% with 95%CI)	89.5 (89.3-89.7)	89.5 (89.2-89.7)	96.7 (96.5-96.9)	89.5 (89.3-89.6)	89.4 (89.2-89.6)	96.7 (96.5-96.9)
AUC (% with 95%CI)		91.7 (88.8-94.7)	87.3 (83.5-91.0)		62.3 (56.6-68)	60.6 (55.2-66)

Abbreviations: PEB, parametric empirical Bayes; MMT, method of mean trends; CA125, cancer antigen 125; AUC, area under roc-curve; CI, confidence interval
* includes borderline, non-epithelial and invasive epithelial
** MMT considered abnormal if >1/2570

The performance for non-epithelial and borderline cancer diagnosed within 1 year of last annual sample is detailed in Supplementary Table 1. Both algorithms performed similarly.

Figure 2 shows the ROC curves for iEOC/PPC diagnosed >1 and $\leq$2 years from last annual sample (secondary analysis). Here MMT (0.613) had slightly higher AUC compared to the CA125 (0.598), although the difference was not significant (p=0.639).

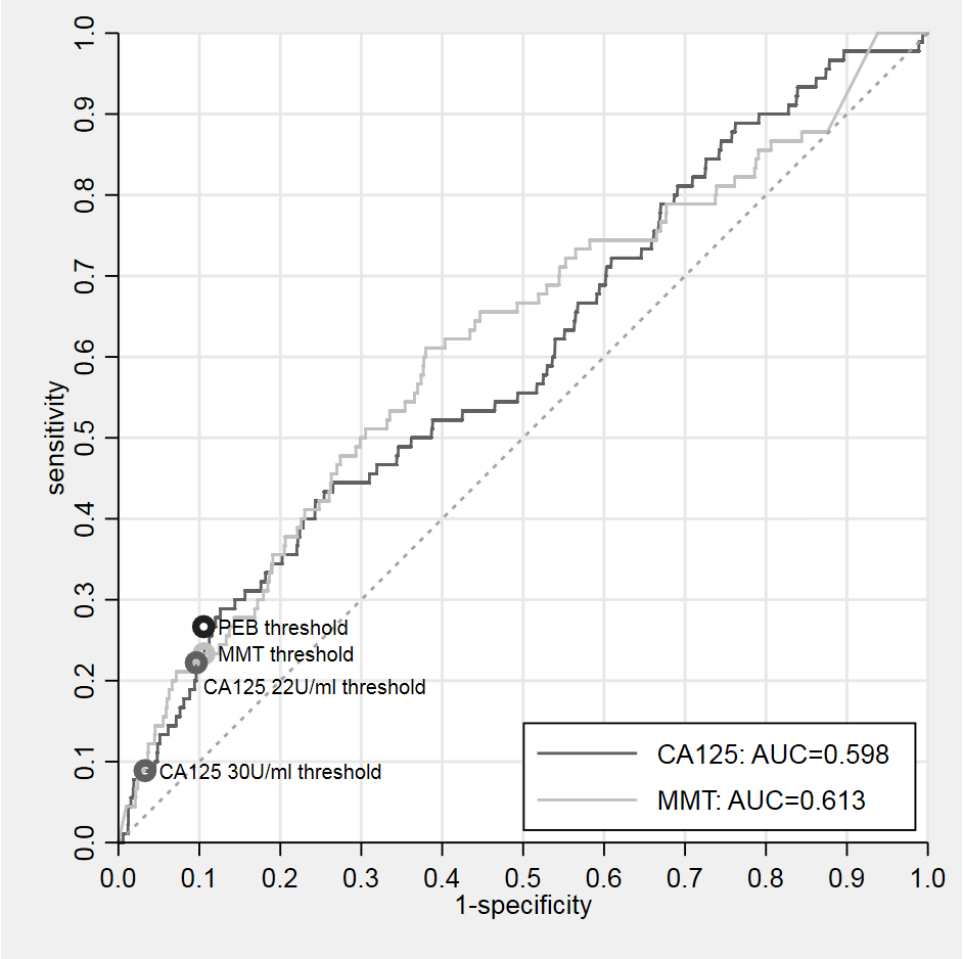


Figure 2 Secondary analysis ROC curves for CA125 interpreted using MMT and threshold rules for detection of iEOC/PPC cases. Circle points on the ROC curves give particular values of sensitivity and specificity provided by MMT and PEB corresponding to cut-offs obtained from the training set (MMT and PEB).

Abbreviations: PEB, parametric empirical Bayes; MMT, method of mean trends; CA125, cancer antigen 125; AUC, area under roc-curve

Discussion

In the largest available serial data set of CA125 results in the general population comprising 347,002 serial annual CA125 measurements from 50,083 women who participated in multimodal screening in UKCTOCS with no selection bias, two serial biomarker algorithms

had high and comparable performance in the context of a first line screening test for invasive epithelial ovarian/tubal/peritoneal cancer and were significantly superior to a CA125 cut-off. We have previously reported that the longitudinal algorithm ROCA outperforms CA125 cut-offs (6). We now show that in comparison to thresholds, other longitudinal algorithms have similar superior performance as a first line test as ROCA (sensitivity 87.1%; specificity 87.6%; AUC 0.915) in UKCTOCS (6). The results emphasize the need to incorporate serial change in biomarker levels in the context of screening and early detection of cancer. Screening is not recommended in the general (low-average risk) population as there is no definitive evidence of a mortality benefit (13). However, our findings have immediate implications for high-risk women in countries where CA125 screening is an option (14). The results clearly show that longitudinal approaches are better tools for the early detection of invasive epithelial ovarian cancer than a single threshold rule which is the current norm.

We compared two serial algorithms, PEB(10) the only other reported serial algorithm that has been used for ovarian cancer screening and our newly developed algorithm (MMT) as a first line test for ovarian cancer screening. The MMT evaluates the dynamics of longitudinal markers by analyzing different trend indices while the PEB models marker trajectory in healthy individuals over time to generate person-specific positivity thresholds. We developed the MMT algorithm and trained it together with the PEB in a random training set which included half the women and half the ovarian/tubal/PPC cancers that were diagnosed prior to 31st Dec 2014 in the MMS arm of UKCTOCS. The ROCA, which is built on a change-point pattern in an individual's CA125 values, was developed on data from previous trials and was prospectively evaluated in UKCTOCS. In future, it will be further refined using the data from the UKCTOCS training set and the refined ROCA will be compared to MMT,

PEB performance in the validation test. The advantages of the MMT and PEB algorithms are that they are computationally simpler than ROCA and therefore can be more easily applied to longitudinal analysis of multiple biomarkers. The MMT algorithm is based on the construction of a logistic regression and therefore for any additional biomarker we only have to calculate trend indices, add them to the logistic regression model and fit it. An advantage of ROCA is that it incorporates tumor doubling into the model, a well-accepted biological dynamic in cancer biology, and is therefore potentially more powerful than algorithms that do not incorporate such biology.

Our results confirm the superiority of serial algorithms for detection of iEOC/PPC diagnosed within one year of the last annual screen. Previous retrospective analysis has involved small sample sets. Drescher et al evaluated PEB in a serial serum CA125 sample set from 44 incident ovarian cancer cases identified from participants in the PLCO (Prostate Lung Colorectal and Ovarian) Cancer Screening Trial Comparison(15). Application of these new algorithms require incorporation into a multimodal strategy with development of cut-offs so that women can be triaged to repeat CA125 testing and second line tests such as transvaginal ultrasound (11) or other novel tests such as circulating tumour DNA (ctDNA) (16). The latter are essential to increase the specificity of the screening strategy and decrease the number of women referred to surgery.

In our secondary analysis where we determined sensitivity of the serial algorithms for detection of cases diagnosed more than one year but within 2 years after the annual sample, both MMT and PEB detected similar small proportions of cases but it is likely that this would not have led to improved lead time. This suggests that further improvements in sensitivity require inclusion of additional ovarian cancer biomarkers to confirm the CA125

trend detected by the serial algorithms. A highly specific marker such as ctDNA would be ideal, but less specific markers such as HE4 may also contribute to earlier diagnosis. We are evaluating the HE4 in this sample set and will report in the near future.

Key strengths of our analysis are the size of the dataset, use of the entire cohort with minimal selection bias, completeness of data on cancers diagnosed in the cohort ensured by linkage to national cancer/death registries using a unique identifier together with two rounds of postal follow-up and independent blinded outcome review of iEOC/PPC. The test results in the validation set were generated by OB who was blinded to outcomes with the unblinding and statistical analysis done independently by MB. The main limitation is that we are only able to assess the algorithms as first line tests. Hence it is not possible to assess the true performance characteristics when incorporated into a multimodal strategy.

In conclusion, our analysis provides definitive evidence of the superiority of longitudinal algorithms compared to single-threshold rules which is the current norm for interpretation of serum CA125 as a first line test in ovarian screening. It is likely that this also applies to other serum markers used in cancer screening. Use of these newer algorithms in ovarian cancer screening requires incorporation into a multimodal strategy and evaluation in clinical trials to assess overall performance**.**

Author contribution

Conception and design: Usha Menon, Oleg Blyuss, Alexey Zaikin, Andy Ryan, Aleksandra Gentry-Maharaj, Ranjit Manchanda

Provision of study data: Andy Ryan, Aleksandra Gentry-Maharaj, Usha Menon, Matthew Burnell, Jatinderpal Kalsi, Ranjit Manchanda, Mahesh Parmar, Steven J. Skates, Ian Jacobs.

Algorithm construction: Oleg Blyuss, Alexey Zaikin, John F. Timms, Inés P. Mariño,

Data analysis and interpretation: Oleg Blyuss, Matthew Burnell, Usha Menon, Andy Ryan, Aleksandra Gentry-Maharaj

Manuscript writing: Oleg Blyuss, Usha Menon, Aleksandra Gentry-Maharaj, Matthew Burnell, Andy Ryan

Final approval of manuscript: All authors

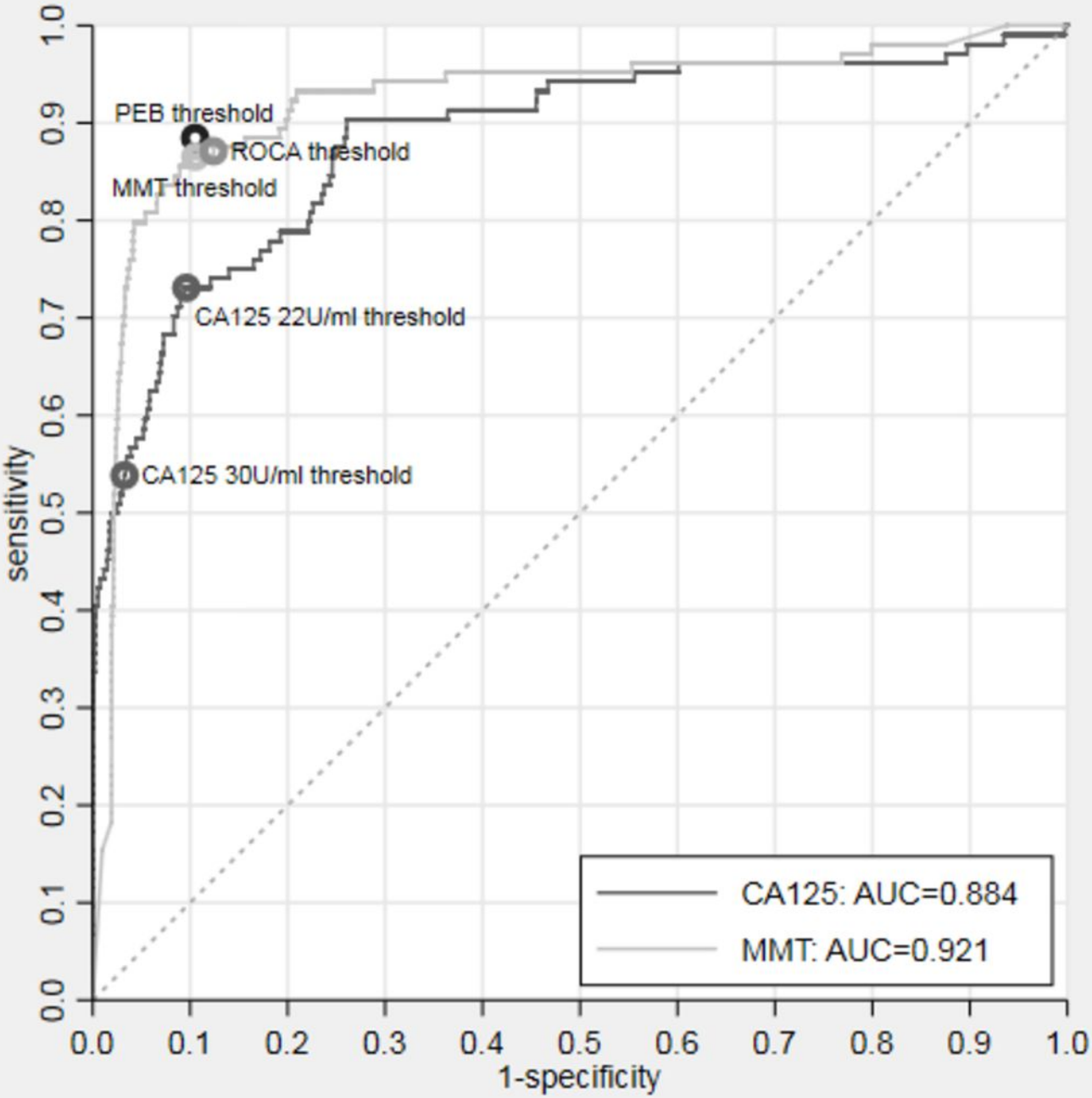
Acknowledgements

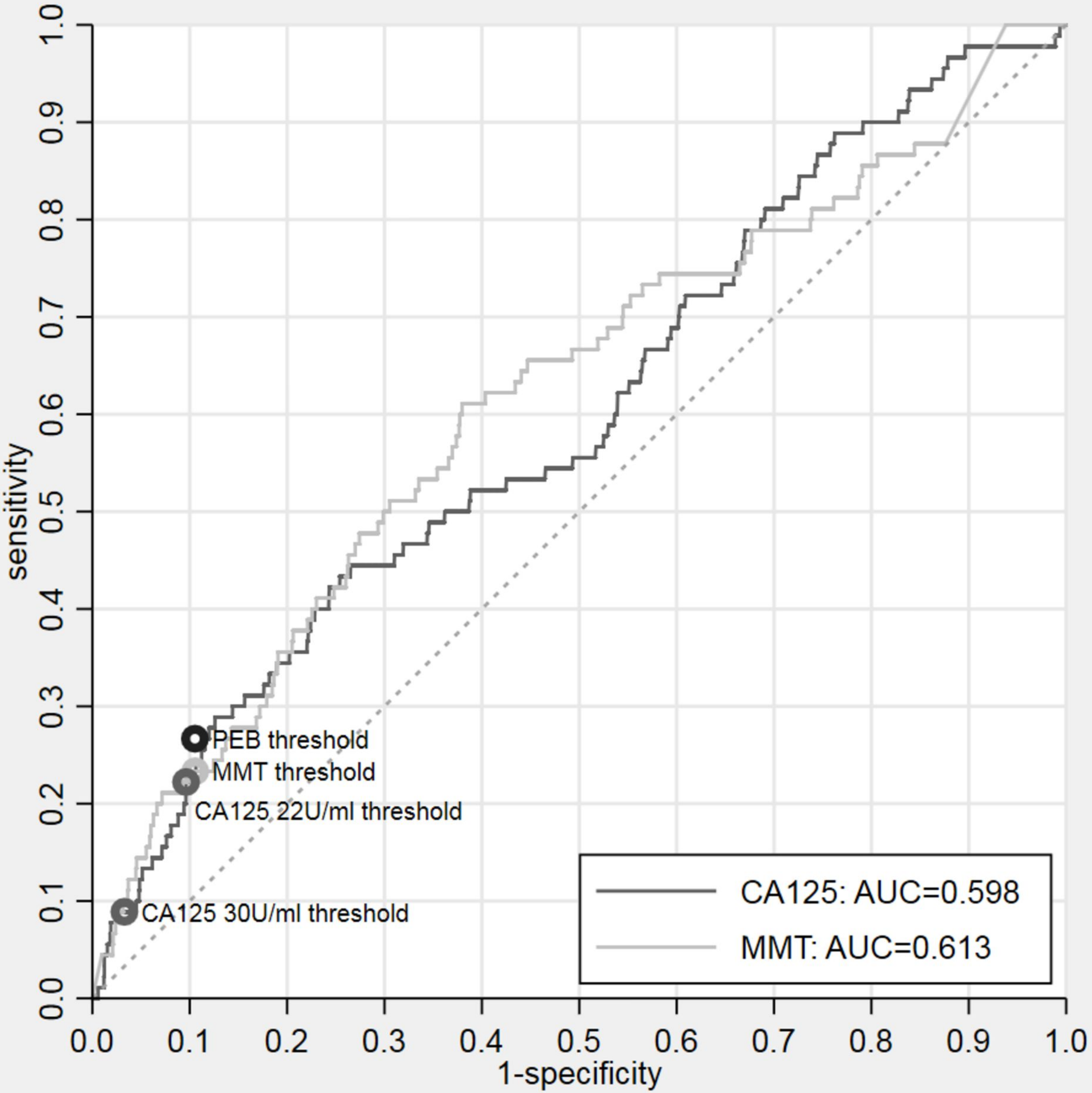
The analysis is part of PROMISE, which was funded through Cancer Research UK PRC Programme Grant A12677 and by The Eve Appeal. It was supported by the National Institute for Health Research (NIHR) University College London Hospitals (UCLH) Biomedical Research Centre. UKCTOCS was core funded by the Medical Research Council, Cancer Research UK, and the Department of Health with additional support from the Eve Appeal, Special Trustees of Bart's and the London, and Special Trustees of UCLH. We thank all the trial participants and all the staff involved in the UKCTOCS trial.

References

1. CRUK. Cancer statistics: Ovarian cancer survival statistics. http://info.cancerresearchuk.org/cancerstats/types/ovary/survival/ 2016. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer/mortality.
2. American Cancer Society: What are the key statistics about ovarian cancer? http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-key-statistics 2016 [cited 2016 07/07/2016]. Available from: http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-key-statistics.

3. Rauh-Hain JA, Krivak TC, Del Carmen MG, Olawaiye AB. Ovarian cancer screening and early detection in the general population. Reviews in obstetrics and gynecology. 2011;4(1):15-21.
4. Cramer DW, Bast RC, Jr., Berg CD, Diamandis EP, Godwin AK, Hartge P, et al. Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. Cancer prevention research (Philadelphia, Pa. 2011;4(3):365-74.
5. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. 2016;387(10022):945-56.
6. Menon U, Ryan A, Kalsi J, Gentry-Maharaj A, Dawnay A, Habib M, et al. Risk Algorithm Using Serial Biomarker Measurements Doubles the Number of Screen-Detected Cancers Compared With a Single-Threshold Rule in the United Kingdom Collaborative Trial of Ovarian Cancer Screening. J Clin Oncol. 2015;33(18):2062-71.
7. Skates SJ, Jacobs IJ, Sjovall K, Einhorn N, Xu FJ, Yu YH, et al. High sensitivity and specificity of screening for ovarian cancer with the risk of ovarian cancer (ROC) algorithm based on rising CA125 levels. Journal of Clinical Oncology. 1996;14(5):2007-.
8. Skates SJ, Pauler DK, Jacobs IJ. Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association*. 2001;96(454):429-39.
9. Marino IP, Blyuss O, Ryan A, Gentry-Maharaj A, Timms JF, Dawnay A, et al. Change-point of multiple biomarkers in women with ovarian cancer. Biomedical Signal Processing and Control. 2017;33:169-77.
10. McIntosh MW, Urban N, Karlan B. Generating longitudinal screening algorithms using novel biomarkers for disease. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2002;11(2):159-66.
11. Menon U, Gentry-Maharaj A, Hallett R, Ryan A, Burnell M, Sharma A, et al. Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Lancet Oncol. 2009;10(4):327-40.
12. Bishop CM. Pattern Recognition and Machine Learning: Springer; 2006. 738 p.
13. Menon U, Karpinskyj C, Gentry-Maharaj A. Ovarian Cancer Prevention and Screening. Obstet Gynecol. 2018.
14. USPSTF. Ovarian Cancer: Screening - Draft Recommendation Statement 2017 [23/04/2018]. Available from: https://www.uspreventiveservicestaskforce.org/Page/Document/draft-recommendation-statement174/ovarian-cancer-screening1#Pod9.
15. Drescher CW, Shah C, Thorpe J, O'Briant K, Anderson GL, Berg CD, et al. Longitudinal screening algorithm that incorporates change over time in CA125 levels identifies ovarian cancer earlier than a single-threshold rule. J Clin Oncol. 2013;31(3):387-92.
16. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359(6378):926-30.


PEB threshold
ROCA threshold
MMT threshold
CA125 22U/ml threshold
CA125 30U/ml threshold
CA125: AUC=0.884
MMT: AUC=0.921
sensitivity
1-specificity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


sensitivity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1-specificity
PEB threshold
MMT threshold
CA125 22U/ml threshold
CA125 30U/ml threshold
CA125: AUC=0.598
MMT: AUC=0.613